Salient Objects in Clutter: Bringing Salient Object Detection to the Foreground
杂乱环境下的显著性物体：将显著性物体检测推向新高度

范登平[1], 程明明[1], 刘姜江[1],
高尚华[1], 侯淇彬[1], and Ali Borji[2]

[1] 计算机学院，南开大学，天津，中国
[2] 计算机视觉研究中心，中佛罗里达大学, 奥尔兰多, 佛罗里达, 美国
http://mmcheng.net/SOCBenchmark/

Abstract. 我们针对显著性物体检测（SOD）模型提出了一种综合评估方案。我们的分析确定了现有SOD数据集存在着严重设计偏差——它们假设每个图像在低杂乱环境中包含至少一个明显突出的显著对象。在现有数据集上进行评估时设计偏差会导致最先进的SOD模型展现出饱和的高性能。然而，当应用于现实世界的日常场景时，这些模型仍然远远不能令人满意。根据我们对显著性数据集的分析,首先确定了一个全面和平衡的数据集应该满足的7个关键点。然后，我们提出一个新的高质量数据集并更新了之前的显著性基准测试。特别地，我们的**SOC**（杂乱环境下的显着对象）数据集包含了来自日常物体类别的显著和非显著物体的图像。除了物体类别的标注之外，每个显著图像都伴随着能够反映现实世界场景中识别挑战性的相关属性。最后，我们在数据集上对各种方法进行基于属性的性能评估。

Keywords: 显著性物体检测· 显著性基准测试· 数据集· 属性

1 简介

本文主要分析了显著性物体检测（SOD）的任务。视觉显著性旨在模仿人类视觉系统选择视觉场景的某个子集的能力。而SOD则侧重于检测场景中吸引最多注意力的物体，然后逐像素地提取物体的轮廓。SOD的优点在于它在许多计算机视觉任务中均有广泛的应用，包括：视觉跟踪 [4]，图像检索 [14,16]，计算机图形学 [9]，内容感知的图像裁剪 [45]和弱监督语义分割 [18,39,40]。

Fig. 1. SOC数据集中的样本图像包括非显著物体图像（第1行）和显著物体图像（第2行到第4 行）。对于显著物体图像，我们提供了实例级真值图（不同颜色表示不同实例）、物体属性和类别标签。有关我们数据集的更多说明，请参阅补充材料。

本文的工作主要受到两个观察的启发。首先，现有的SOD数据集 [2, 5, 10, 11, 23, 26, 29, 32, 43, 44] 在数据收集过程或数据质量方面存在缺陷。具体而言，大多数数据集假设图像包含至少一个显著物体，因此它们丢弃了不包含显著物体的图像。我们将此称为*数据选择偏差*。此外，现有数据集主要包含具有单个物体的图像或处于低杂乱环境中的多个物体（通常是人）。这些数据集不能充分反映现实世界中图像的复杂性，在现实世界中，场景通常包含多个杂乱的物体。这样导致的结果便是在现有数据集上训练的表现最佳的模型几乎已经可以使性能达到饱和（例如，在大多数数据集上，$F\text{-}measure > 0.9$ ）但它们在现实场景上的表现却无法令人满意（例如，Table 3中$F\text{-}measure < 0.45$ ）。这是因为在之前数据集上训练出来的模型更加偏向较为理想的场景，所以一旦它们应用于现实世界中的场景时，其有效性可能会受到极大削弱。因此，为了解决此问题，有必要构建更接近实际条件的数据集。

其次，在现在的数据集上我们只能分析模型的整体性能，这些数据集都缺乏反映现实场景中所面临的挑战的各种属性。因此，引入这些属性有助于1）更深入地了解SOD问题，2）研究SOD模型的优缺点，3）对于不同的应用来说，从不同的角度客观地评价模型的性能，其评价结果可能是不同的。

考虑到上述两个问题，我们做了2个贡献。我们的主要贡献是构建了一个新的高质量SOD数据集，将其命名为**SOC**（Salient Objects in Clutter）。迄今为

止，SOC是最大的实例级SOD数据集，它包含来自80多个常见类别的6,000张图像。它与现有数据集的不同之处在于三个方面：1）显著物体具有类别注释，可用于诸如弱监督SOD任务之类的新型研究，2）包含非显著图像，使该数据集更接近真实世界场景，并且比现有数据集更具挑战性，3）显著物体具有反映真实世界中面临的特定情况的属性，例如运动造成的模糊，遮挡和杂乱的背景。因此，我们的SOC数据集缩小了现有数据集与现实世界场景之间的差异，并提供了更合理的基准测试（见 Fig. 1）。

此外，我们针对几种最先进的卷积神经网络（CNN）模型进行了综合评估 [8, 15, 17, 23, 24, 28, 31, 36, 38, 48–51]。为了评估模型性能，我们引入了三个评估指标来度量检测结果的区域相似性、分割的像素精度以及结果的结构相似性。此外，我们提供基于属性的性能评估。这些属性允许我们更深入地理解模型并且进一步指出了具有潜力的研究方向。我们相信，该数据集和基准测试会对未来的SOD研究，特别是对于面向应用的模型开发产生非常大的影响。整个数据集和分析工具将免费向公众发布。

2 相关工作

在本节中，我们将简要讨论那些现有的、为SOD任务设计的数据集，尤其从标注类型、每个图像的显著物体数量、图像数量和图像质量等方面展开。我们也将回顾基于卷积神经网络的SOD模型。

(a) 原图 (b) 逐像素 (c) 实例 (d) 分割

Fig. 2. 之前的SOD数据集仅通过绘制（b）显著物体的逐像素轮廓来标注图像。不同于来自（d）MSCOCO物体分割数据集 [27]（该数据集中物体不一定是显著的），本文的工作焦点在于对（c）显著物体的实例级的分割。

2.1 数据集

早期数据集要么局限于图像的数量，要么受限于显著物体的标注质量。例如，数据集**MSRA-A** [29]和**MSRA-B** [29]中的显著物体基本是以标定框的形式进行标注。**ASD** [1]和**MSRA10K** [11]在每张图像中大多只包含一个显著物

体，而**SED2** [2] 数据集在单张图像中包含两个物体但仅包含100个图像。为了提高数据集的质量，近年来研究人员开始收集具有相对复杂和杂乱背景具有多个物体的数据集。这些数据集包括**DUT-OMRON** [44]、**ECSSD** [43]、**Judd-A** [5] 和**PASCAL-S** [26]。与之前的数据集相比，这些数据集在标注质量和图像数量方面得到了改进。数据集**HKU-IS** [23]，**XPIE** [41]和**DUTS** [37]通过收集具有多个显著物体的大量逐像素标注图像（ Fig. 2 （b））来克服这些缺点。然而，他们忽略了非显著物体，并且没有提供实例级（Fig. 2（c））的显著物体标注。除此之外， [19]的研究人员收集了大约6k张简单的背景图像（大多数是纯纹理图像）来表示非显著的场景，但由于真实场景更复杂，因此该数据集不足以反映真实场景。**ILSO** [22]数据集包含实例级显著对象标注，但其标注的边界仍较为粗糙，如Fig. 5（a）所示。

总而言之，现有数据集主要集中在具有简单背景的清晰显著物体的图像上。考虑到现有数据集的上述局限性，需要一个包含具有非显著物体、“户外场景”的纹理以及具有属性的显著物体的更贴近真实场景的数据集以用于该领域的未来研究。这样的数据集可以帮助研究者深入洞察SOD模型的弱点和优势。

2.2 模型

我们根据任务的数量来对最先进的SOD深度神经网络模型进行分类。

单任务模型的唯一目标是检测图像中的显著物体。在**LEGS** [36]中，局部信息和全局对比度分别由两个不同的深度卷积神经网络捕捉，然后将它们融合以生成显著图。在文献 [51]中，Zhao等人为SOD提出了一个多上下文深度学习框架（**MC**）。Li等人 [23]（**MDF**）提出使用从深度卷积神经网络中提取多尺度特征来导出显著图。Li等人 [24]提出了一个深度对比网络（**DCL**），它不仅考虑了像素信息，还将分割级别的引导融合到网络中。Lee等人 [15]（**ELD**）考虑了从卷积神经网络中提取的高级特征和手工设计的特征。Liu等人 [28]（**DHS**）设计了一个两阶段的网络，其中一个产生了一个粗略的缩减预测图，然后是另一个网络对预测图的细节进行细化并分层和逐步地对预测图进行上采样。Long等人 [30]提出了一种全卷积网络（**FCN**），该网络使密集像素预测问题的端到端训练变得可行。**RFCN** [38]使用了一个重复的全卷积网络将粗糙的预测图作为显著性的先验信息，并以一种逐阶段的方式改进了最后生成的预测图。**DISC** [8]框架被提出用于细粒度图像的显著性计算。利用两个堆叠的卷积神经网络来分别获得粗糙和细粒度的显著图。**IMC** [48]通过全卷积网络在不同层面上整合了显著性线索。它是一种可以有效地利用学习到的语义线索和高阶区域统计数据来获得精确边缘的SOD模型。最近，业内提出了一种具有短连接（**DSS**）的深层架构 [17]。Hou等人添加了从高级别特征到基于HED [42]架

Table 1. 基于卷积神经网络的SOD模型。我们将这些模型分为单任务（S-T）和多任务（M-T）***Training Set***：MB是MSRA-B数据集 [29]。MK是MSRA-10K [11] 数据集。ImageNet数据集是指 [34]。D是DUT-OMRON [44]数据集。H是HKU-IS [23]数据集。P是PASCAL-S [26]数据集。P2010是PASCAL VOC 2010语义分割数据集 [12]。***Base Model***：VGGNet，ResNet-101，AlexNet，GoogleNet是所基于的模型。***FCN***：模型是否使用全卷积网络。***Sp***：模型是否使用超像素。***Proposal***：模型是否使用目标提取。***Edge***：模型是否使用边缘或轮廓信息

	No	Model	Year	Pub	#Training	*Training Set*	*Base Model*	*FCN*	*Sp*	*Proposal*	*Edge*
S-T	1	**LEGS** [36]	2015	CVPR	3,340	MB + P	—	×	×	✓	×
	2	**MC** [51]	2015	CVPR	8,000	MK	GoogLeNet	×	✓	×	×
	3	**MDF** [23]	2015	CVPR	2,500	MB	—	×	✓	×	✓
	4	**DCL** [24]	2016	CVPR	2,500	MB	VGGNet	✓	✓	×	×
	5	**ELD** [15]	2016	CVPR	9,000	MK	VGGNet	×	✓	×	×
	6	**DHS** [28]	2016	CVPR	9,500	MK+D	VGGNet	×	×	×	×
	7	**RFCN** [38]	2016	ECCV	10,103	P2010	—	✓	✓	×	✓
	8	**DISC** [8]	2016	TNNLS	9,000	MK	—	×	✓	×	×
	9	**IMC** [48]	2017	WACV	6,000	MK	ResNet-101	✓	✓	×	×
	10	**DSS** [17]	2017	CVPR	2,500	MB	VGGNet	✓	×	×	✓
	11	**NLDF** [31]	2017	CVPR	2,500	MB	VGGNet	✓	×	×	×
	12	**AMU** [49]	2017	ICCV	10,000	MK	VGGNet	✓	×	×	✓
	13	**UCF** [50]	2017	ICCV	10,000	MK	—	✓	×	×	×
M-T	1	**DS** [25]	2016	TIP	10,000	MK	VGGNet	✓	✓	×	×
	2	**WSS** [37]	2017	CVPR	456K	ImageNet	VGGNet	✓	✓	×	×
	3	**MSR** [22]	2017	CVPR	5,000	MB + H	VGGNet	✓	×	✓	✓

构的低级别特征的连接，并且该架构实现了良好的性能。**NLDF** [31]整合了局部和全局特征，并在标准交叉熵损失中增加了边界损失项以训练端到端网络。**AMU** [49]是一个通用的聚合多级卷积特征的框架。它将粗略的语义和详细的特征映射集成到多个分辨率中。然后，它自适应地学习如何将每个分辨率的特征图和预测的显著图与组合特征图相结合。**UCF** [50]被提出的目的是提高显著性检测的鲁棒性和准确性。他们在特定的卷积层之后引入了重新形成的丢弃层，以构建一个不确定的内部特征单元集合。此外，他们在有效的混合上采样方法之后提出了重新计算的丢弃层，以减少解码器网络中反卷积操作的棋盘伪像。

多任务模型目前包括三种方法，**DS**，**WSS**和**MSR**。**DS** [25]模型建立了一个多任务的学习方案，该模型被用于探索显著性检测和语义图像分割之间的内在相关性，它们共享全卷积网络层中的信息以生成物体感知的有效特征。最近，Wang等人 [37]提出了一个名为**WSS**的模型，该模型开发了一种使用图像级标签进行显著性检测的弱监督学习方法。首先，他们共同训练前景推理网络（FIN）和全卷积网络进行图像分类。然后，他们使用前景推理网络迭代CRF来加强空间标签一致进而预测显著图。**MSR** [22]与多尺度组合聚类和基于MAP的 [47]子集优化框架相结合后被用于显著区域检测和显著物体轮廓检测，该模型使用三个已知的具有共享参数的VGG网络流和用于在不同尺度上融合结果的学习注意力模型，使得研究者能够获得良好的预测结果。

我们根据提出的SOC数据集对大量最先进的基于卷积神经网络的模型（见Table 1）进行了基准测试，发现了当前模型存在的问题并指出了未来的研究方向。

3 提出的数据集

在本节中，我们将介绍新的旨在详细反映真实世界场景的具有挑战性的SOC数据集。来自SOC的样例图像如Fig. 1所示。此外，关于SOC的类别和属性的统计分别如 Fig. 4和Fig. 6所示。基于现有数据集的优点和缺点，我们确定了全面和平衡的数据集应该满足的七个关键方面。

1) 非显著物体的存在。几乎所有的现有SOD数据集都假设图像包含至少一个显著物体并丢弃了不包含显著物体的图像。但是，这种假设是导致*数据选择偏差*的过于理想化的设定。在真实场景的设定中，图像并不总是包含显著物体。例如，一些无定形的背景图像，如天空，草和纹理根本不包含显着的物体 [6]。非显著物体或背景“元素”可能占据整个场景，因此严重限制了显著物体的可能位置。Xia等人 [41]通过判断什么是显著物体和什么不是显著物体，提出了先进的SOD模型，说明非显著物体对推理显著物体至关重要。这表明非显著物体应该和SOD中的显著物体受到同等重视。包含一定数量的非显著物体图像会使得数据集更接近真实场景，同时也使得SOD任务变得更具挑战性。因此，我们将“非显著物体”定义为“没有显著物体的图像或具有“元素”性质的图像。如 [6, 41]中所述，“元素”类别包括（a）密集分布的相似物体，（b）形状模糊，和（c）没有语义的区域，分别如Fig. 3 （a）- （c）所示。

(a) (b) (c)

Fig. 3. 一些非显著图像的示例

基于非显著物体的定义，我们从DTD [21]数据集中收集了783个纹理图像。为了丰富多样性，从互联网和其他数据集中收集了2217幅图像，包括极光，天

空，人群，商店和许多其他类型的真实场景 [26, 27, 32, 35]。我们相信，纳入足够的非显著物体会为未来的研究工作开辟了一个有希望的方向。

2) 图像的数量和类别。相当数量的图像对于捕捉现实世界场景的多样性和丰富性至关重要。此外，大量的数据可以让SOD模型避免过拟合并增强泛化能力。为此，我们收集了来自80多个类别的6,000张图像，其中包含3,000张带有显著物体的图像和3,000张没有显著物体的图像。我们将数据集分为训练集，验证集和测试集，比例为6：2：2。为了确保公平性，测试集不会发布，而是通过我们的网站提供在线测试[3]。Fig. 4 （a）显示了每个类别的显著物体的数量。它表明“人”类别占很大比例，这是合理的，因为人们通常与其他对象一起出现在日常场景中。

3) 显著物体的全局/局部颜色对比。如 [26]中所述，术语“显著”与前景和背景的全局/局部对比度有关。因此检查显著物体是否易于检测是非常重要的。对于每个物体，我们分别计算前景和背景的RGB颜色直方图。然后，利用χ^2距离来测量两个直方图之间的距离。全局和局部颜色对比度分布分别如Fig. 4（b）和（c）所示。与ILSO相比，我们的SOC中具有低全局颜色对比度和局部颜色对比度的物体占据更大的比例。

4) 显著物体的位置。中心偏差被认为是显著性检测数据集中影响最大的偏差之一 [3, 20, 26]。Fig. 4 （d）展示出了一组图像及其叠加图。可以看出，虽然显著的物体位于不同的位置，但是叠加图仍然表明这组图像是存在中心偏置的。以前的基准测试通常采用这种不准确的方式来分析显著物体的位置分布。为了避免这种误导现象，我们绘制了Fig. 4（e）中两个量r_o和r_m的统计情况，其中r_o 和r_m 分别表示物体中心和物体中最远（边缘）点离图像中心有多远。将r_o和r_m除以图像对角线长度的一半以进行归一化，使得$r_o, r_m \in [0, 1]$。从这些统计数据中，我们可以观察到数据集中的显著物体不受中心偏差的影响。

5) 显著物体的大小。每个显著物体实例的大小被定义为物体面积占图像总面积的比例 [26]。如Fig. 4（g）所示，与仅有的实例级ILSO数据集 [22]相比，SOC中的显著物体的大小的变化范围更广泛。此外，SOC中的中型物体具有更高的比例。

6) 高质量的显着对象标签。正如 [17]中所注意到的，在ECSSD数据集（具有1,000个图像）上的训练允许模型获得比其他数据集（例如，MSRA10K，具有10,000个图像）获得更好的结果。这表明除了规模之外，数据集质量也是一个重要因素。为了获得大量高质量的图像，我们从MSCOCO数据集 [27]中随机选择图像，这是一个大型的真实世界数据集，其中的物体用多边形标注（例如，粗略标注）。高质量标注在提高SOD模型的准确性方面也起着关键作用 [1]。为

[3] http://dpfan.net/SOCBenchmark/

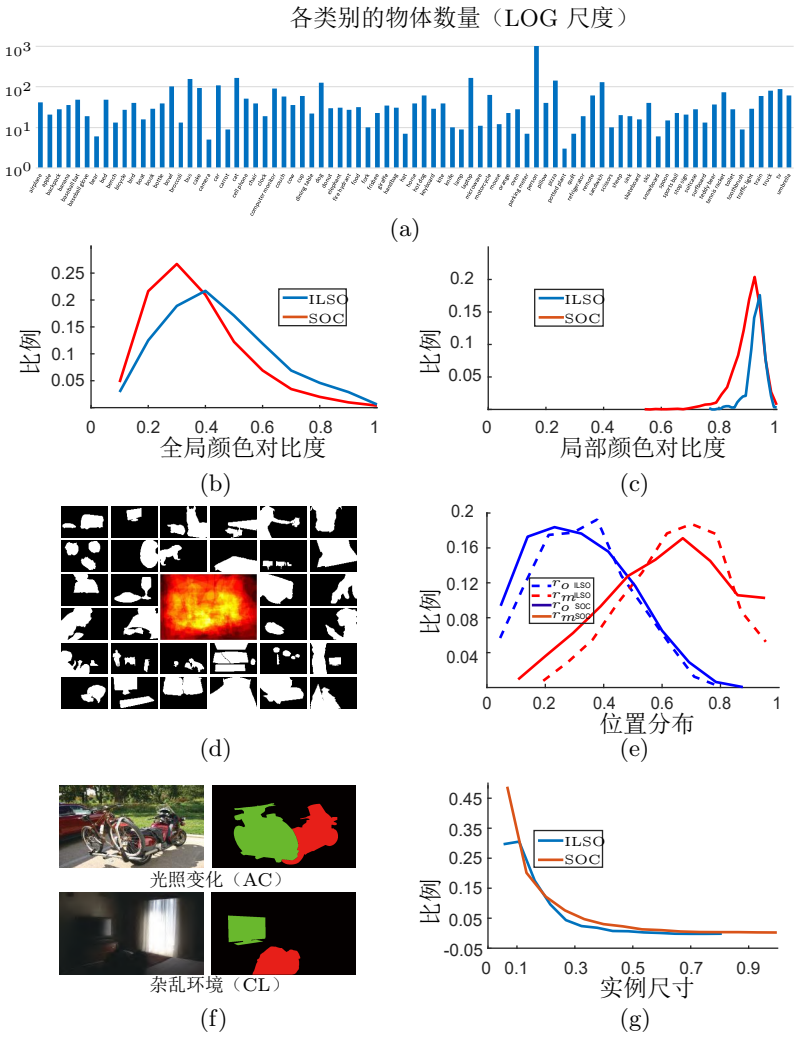
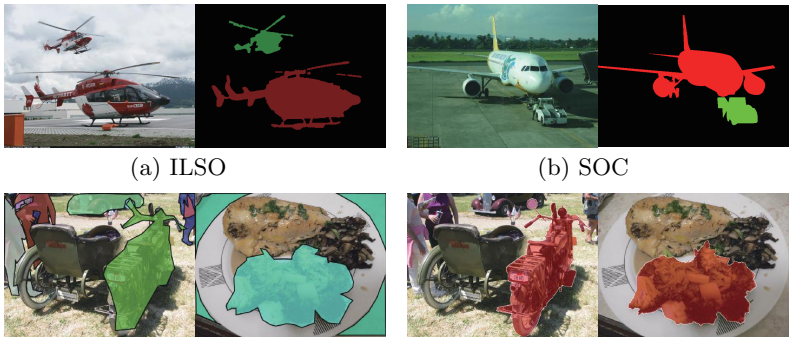


Fig. 4. （a）我们的SOC数据集中每个类别标注的实例的数量。（b，c）全局颜色对比度和局部颜色对比度的统计数据。（d）来自我们的数据集及其叠加图的一组显著图。（e）SOC中的显著物体的位置分布。（f）不同属性的视觉例子。（g）SOC和ILSO [22]的实例大小分布。

此，我们使用逐像素的标注来重新标记数据集。类似于著名的SOD任务导向基准测试数据集 [1,2,11,19,22,23,29,32,37,41,43]，我们没有使用眼动仪设备。我们采取了多个步骤来提供高质量的注释。这些步骤包括两个阶段：（i）我们要求5个观众使用标定框标记他们认为在每个图像中较为显著的物体。（ii）保留大多数（≥ 3）观众在显著性上意见相同的物体（标定框的IOU>0.8）。在第一



(a) ILSO (b) SOC

(c) MSCOCO (d) SOC

Fig. 5. 与最近提出的数据集的比较（a）实例级数据集ILSO [22]（用不连续的粗略边界标注），（c）MSCOCO数据集 [27]（用多边形标注），我们的（b，d）SOC数据集中的标注边界更平滑，质量更高。

Table 2. 显著物体图像的属性列表和相应描述。通过观察现有数据集的特征，我们总结了这些属性。一些视觉示例可以在Fig. 1和Fig. 4（f）中找到。有关更多示例，请参阅补充材料。

属性描述	
AC	**光照变化** 物体区域中明显的光照变化。
BO	**大物体** 物体面积和图像面积的比值大于0.5
CL	**杂乱** 物体周围的前景和背景区域具有相似的颜色，我们将全局颜色对比度值大于0.2局部颜色对比度值小于0.9的图像标记为杂乱图像。
HO	**异构物体** 由视觉上独特/不相似的部分组成的物体。
MB	**运动模糊** 由于相机或运动的抖动使得物体具有模糊的边界。
OC	**遮挡** 物体被部分或全部遮挡
OV	**超出视野** 物体的部分区域超出了图像边界。
SC	**形状复杂性** 物体有比如纤细组件之类的复杂的边界，比如动物的脚和洞
SO	**小物体** 物体面积和图像面积的比率小于0.1。

阶段之后，我们有3,000个用标定框标注的显著物体图像。在第二阶段，我们根据标定框的提示进一步手工标记显著物体的逐像素轮廓。请注意，我们有10名志愿者参与了整个步骤以交叉检查标注的质量。最后，我们保留了3,000张具有高质量的实例级标记显著物体的图像。如Fig. 5（b，d）所示，我们的物体边界的标注是精确、清晰和平滑的。在标注过程中，我们还添加了一些未在MSCOCO数据集中标记的新类别（例如，计算机显示器，帽子，枕头）[27]

7) 具有属性的显着对象。在数据集中图像的属性信息有助于研究者客观评估模型在不同类型的参数上的性能。它还允许对模型失败情况的检查。为此，我们定义了一组属性来表示真实场景中面临的特定情况，例如运动模糊，遮挡

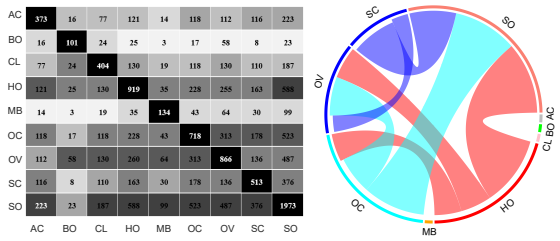


Fig. 6. 左：SOC数据集中显著图像的属性分布。网格中的每个数字表示图像的出现次数。右：基于出现频率绘制的属性之间的主要依赖关系。连接的较大宽度表示属性对其它属性的依赖较高。

和杂乱的背景（Table 2中总结）。请注意，因为这些属性不是独占的，所以一个图像可以使用多个属性进行标注，。

受 [33]的启发，如Fig. 6左边所示，我们展现了数据集图片属性的分布情况。*SO*类型具有最大比例是因为精确的实例级（例如，Fig. 2中的网球拍）的标注。因为现实世界的场景由不同视觉特色的材料组成，所以*HO*类型占很大比例。*MB* 类型在视频帧中比静态图像更常见，但有时也会出现在静态图像中。因此，*MB*类型在我们的数据集中占相对较小的比例。由于真实图像通常包含多个属性，为此我们在Fig. 6右侧根据出现的频率显示了属性之间的主要依赖关系。例如，包含许多异构物体的场景可能具有大量彼此阻挡并形成复杂空间结构的物体。因此，*HO*类型与*OC*类型，*OV*类型和*SO* 类型具有强依赖性。

4 基准测试模型

在本节中，我们在SOC数据集上呈现了16个SOD模型的评估结果。几乎所有基于卷积神经网络的具有代表性的SOD模型都进行了评估。但是，由于某些模型的代码不公开，因此我们对此类模型不予考虑。此外，大多数模型都没有针对非显著物体检测进行优化。因此，公平起见，我们只使用SOC数据集的测试集来评估SOD模型。我们在Sec. 4.1中描述了评估指标。SOC数据集的整体模型性能见Sec. 4.2和Table 3 中，而针对各个属性的性能评估结果（例如，光照变化属性上的性能表现）见Sec. 4.3和Table 4 中。我们公开了评估脚本并且在网站上会提供在线评估测试。

4.1 评估指标

在强监督评估框架中，给定由SOD模型生成的预测图M及真值掩膜G，我们寄希望于评估指标告诉我们究竟是哪一种模型能够生成最佳结果。在这里，我们在SOC数据集上使用三种不同的评估指标来评估SOD模型。

逐像素精度ε 。区域相似性评估方法不考虑真负类的显著性分布。作为补补充，我们也计算M和G之间的归一化（[0,1]）后的平均绝对误差（MAE），将其定义为：

$$\varepsilon = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} ||M(x,y) - G(x,y)||, \tag{1}$$

其中W和H分别是图像的宽度和高度。

区域相似性F 。为了测量两张图片的各区域相匹配的程度，我们使用F-$measure$，该方法定义如下：

$$F = \frac{(1+\beta^2) Precision \times Recall}{\beta^2 Precision + Recall}, \tag{2}$$

其中$\beta^2 = 0.3$由 [1]提出并用于平衡召回率和精度。然而，在计算召回率和精度时，F-$measure$并没有对黑色（全零矩阵）真值图进行定义。在这种情况下，不同的前景图会得到相同的结果（0），而这显然是不合理的。从而可以得出结论，F-$measure$不适合于评估非显著物体检测的结果。

综上，ε和F这两个指标都是基于逐像素的计算方式，因此经常忽略结构相似性。行为视觉研究表明人类视觉系统对场景结构非常敏感 [13]。在许多应用中，我们期望SOD模型的检测结果能够保留物体的结构信息。

结构相似性S 。Fan等人 [13] 提出的S-$measure$同时考虑区域和物体两个层次上的相似性来评估检测结果的结构相似性。因此，我们也使用S-$measure$来评估M和G 之间的结构相似性。需要特别注意的是我们对接下来整体性能表现的评估和分析是基于S-$measure$进行的。

4.2 指标统计

为了获得整体结果，我们对评估指标的分数取平均值。设$\eta \in \{F, \varepsilon, S\}$，度量平均得分的公式为：

$$M_\eta(D) = \frac{1}{|D|} \sum_{I \in D} \bar{\eta}(I_i), \tag{3}$$

其中$\bar{\eta}(I_i)$是模型在图像数据集D内图像I_i上的评估得分。

Table 3. 三种指标下SOD模型的性能。F代表区域相似性，ε是平均绝对误差，S 是结构相似性。↑代表数字越高越好，反之亦然。评估结果根据式(3)在SOC数据集上通过计算得出。$S_{all}, F_{all}, \varepsilon_{all}$ 分别表示用S, F, ε指标来表示的整体性能表现。**加粗**表示最好成绩。

	单任务													多任务		
指标	LEGS	MC	MDF	DCL	AMU	RFCN	DHS	ELD	DISC	IMC	UCF	DSS	NLDF	DS	WSS	MSR
	[36]	[51]	[23]	[24]	[24]	[38]	[28]	[15]	[8]	[48]	[50]	[17]	[31]	[25]	[37]	[22]
F_{all} ↑	.276	.291	.307	.339	.341	**.435**	.360	.317	.288	.352	.333	.341	.352	.347	.327	**.380**
S_{all} ↑	.677	.757	.736	.771	.737	.814	.804	.776	.737	.664	.657	.807	**.818**	.779	.785	**.819**
ε_{all} ↓	.230	.138	.150	.157	.185	.113	.118	.135	.173	.269	.282	.111	**.104**	.155	.133	.113

单任务：对于单任务模型，在整个SOC数据集上性能表现（Table 3中的S_{all}）最佳的模型是NLDF [31] ($M_S = 0.818$)，其次是RFCN [38] ($M_S = 0.814$)。MDF [23]和AMU [49]使用边缘线索来提升显著图提取的准确度但却未能达到理想的目标。为了使用图像的局部区域信息，MC [51]，MDF [23]，ELD [15]和DISC [8]尝试使用超像素方法将图像分割成数个区域，然后从这些区域中提取特征，但这是较为复杂而耗时的。为了进一步提高性能，UCF [50]，DSS [17]，NLDF [31]和AMU [49]利用全卷积网络来改善SOD模型的性能（Table 4中的（S_{sal}））。其他一些方法诸如DCL [24]和IMC [48] 则尝试将超像素与全卷积网络结合起来构建一个强大的模型。此外，RFCN [38]将包括边缘和超像素的两个相关线索组合到全卷积网络中进而在整个数据集上获得了良好的性能（$M_F = 0.435$, $M_S = 0.814$）。

多任务：与上述模型不同，MSR [22]使用三个密切相关的步骤去检测实例级显著物体：估计显著图，检测显著物体轮廓，以及识别显著物体的实例。它创建了一个多尺度显著性检测网络，可以实现最高性能（S_{all}）。其他两个多任务模型DS [25]和WSS [37]同时利用分割和分类结果生成显著图从而获得适度的性能提升。值得一提的是，尽管WSS是一种弱监督的多任务模型，但它仍然可以实现与其他全监督的单任务模型相当的性能。因此，基于弱监督和多任务的模型可能是未来的研究方向。

4.3 基于属性的评估

如Sec. 3和Table 2所示，我们为显著图像分配了属性。每个属性代表现实世界场景的显著性检测中存在的挑战性问题。这些属性允许我们识别具有主导性特征（例如，杂乱环境的存在）的图像集合，它们对于SOD模型的性能解释以及如何将SOD与面向应用的任务相关联是非常重要的。例如，sketch2photo应

Table 4. 在SOC显著性物体子数据集上基于属性的性能表现。对于每一个模型，分数对应于在特定属性的所有测试图像上的结构相似性M_S（见Sec. 4.1）的平均值，分数越高性能表现越好，**加粗**表示最高成绩，平均显著物体检测性能S_{sal}在第一行通过结构相似性S呈现，$^+$和$^-$分别表示与平均值相比之下的性能增加和减少。

	单任务													多任务		
属性	LEGS [36]	MC [51]	MDF [23]	DCL [24]	AMU [24]	RFCN [38]	DHS [28]	ELD [15]	DISC [8]	IMC [48]	UCF [50]	DSS [17]	NLDF [31]	DS [25]	WSS [37]	MSR [22]
S_{sal}	.607	.619	.610	.705	.705	.709	**.728**	.664	.629	.679	.678	.698	.714	.719	.676	**.748**
AC	.625	.631	.614	.734	.736	.744	**.745**	.673	.644	.702	.714	.726	.737	.764	.691	**.789**
BO	.509	.490	.461$^-$	.610	.569	.540	.590	.576	.517	**.701**$^+$	.636	.496$^-$	.568	.685	.566	.667
CL	.620	.635	.566	.699	.708	.714	**.743**	.658	.635	.696	.704	.677$^-$	.713	.729	.678	**.756**
HO	.666	.666	.648	.745	.755	.759	**.766**	.706	.681	.715	.744	.748	.755	.756	.707	**.777**
MB	.543$^-$	.603	.615	.693	.706	.715	**.722**	.639	.600	.689	.682	.695	.685	.711	.641	**.757**
OC	.609	.617	.608	.708$^+$	**.725**$^+$	.711	.716	.658	.630	.672	.701$^+$	.689	.709	.725$^+$	.672	**.740**
OV	.548	.584	.568	.699	**.708**$^+$	.687	.706	.637	.573	.693$^+$	.685$^+$	.665	.688	.722$^+$	.624	**.743**
SC	.608	.620	.669$^+$	.738	.731	.735	**.763**	.688	.653	.690	.722$^+$	.746$^+$	.745	.724	.677	**.773**
SO	.573$^-$	.601	.621	.691	.685	.698	**.713**	.644	.614	.648$^-$	.650	.696$^-$	.703	.696	.659	**.730**

用 [7]更喜欢在大物体上具有良好性能的模型，而这可以通过基于属性的性能评估方法来辨别。

结果在Table 4中，我们显示了各种SOD模型在特定属性表征的数据集子集上的性能。由于篇幅限制，在以下部分中，我们仅选择一些代表性属性进行进一步分析。更多细节可以在补充材料中找到。

大物体（BO） 当物体与相机距离很近时，经常会出现大物体（*BO*）场景，因此在图片中可以清楚地看到微小的文字或图案。然而在这种情况下，倾向于关注局部信息的模型将被严重误导，导致较大的性能损失（例如，DSS [17]损失了28.9%的性能，MC [51] 损失了20.8%的性能以及RFCN [38]损失了23.8%的性能）。然而，IMC [48] 模型的性能表现略微上升了3.2%。在深入了解该模型的流程后，我们得出了一个合理的解释，即IMC使用粗略预测图来表达语义，并利用过度分割的图像来补充结构信息，从而在*BO*类型图像上获得了令人满意的结果。但是，过度分割的图像无法弥补缺失的细节，因此会导致此类模型在*SO*类型图像上的性能下降4.6%。

小物体（SO） 对于所有SOD模型来说，识别*SO*类型都是一个较为困难的问题。在此类图像上所有模型都遇到了性能下降（例如，从DSS [17]的下降0.3%到LEGS [36]的下降5.6%），因为在卷积神经网络的下采样期间很容易忽视小物体。DSS [17]是唯一一个在*SO*类型图像上性能仅有略微下降的模型，而它在*BO*类型图像上的性能损失最大（28.9%）。MDF [23]使用多尺度超像素图像作为网络的输入，因此它能够很好地保留了小物体的细节。然而，由于超像

素的大小有限，MDF仍无法有效地感知全局语义，导致在*BO*类型图像上出现大的识别失败概率。

遮挡（*OC*）在遮挡场景中，物体被部分遮挡。因此，SOD模型需要捕获全局语义以弥补物体信息的不完整。为此，DS [25]和AMU [49] 利用下采样过程中的多尺度特征生成融合显著图；UCF [50]提出了一种不确定的学习机制来学习不确定的卷积特征。所有这些方法都试图获得包含全局和局部特征的显著图。不出所料，这些方法在*OC*类型上取得了相当不错的效果。基于上述分析，我们还发现这三个模型在需要更多语义信息的场景上的性能表现非常好，如*AC*，*OV*和*CL*类型。

异构物体（*HO*）类型场景在现实生活中很常见。在*HO*类型的图像上不同模型的性能分别比其在总数据集上的平均性能有所提升，基本处于3.9 % 至9.7 %之间的波动上。我们怀疑这是因为*HO*类型占总数据集的比例较大，客观上使SOD模型更适合这个属性。这一结果在某种程度上证实了我们在Fig. 6中的统计数据。

5 讨论和结论

据我们所知，这项工作提出了目前最大尺度的针对基于卷积神经网络的显著性物体检测模型的性能评估方案。我们的分析指出了现有SOD数据集中存在的严重的*数据选择偏差*。这种设计偏差导致最先进的SOD算法在现有数据集上进行评估时几乎达到了几近饱和的高性能表现，但在应用于现实世界日常场景时仍然远远不能令人满意。基于我们的分析，我们首先确定了全面和平衡的数据集应该满足的7 个重要方面。我们首先构建了高质量的SOD数据集**SOC**。它包含来自日常生活的自然环境中的更接近真实环境的显著物体图像。SOC数据集将随着时间的推移而发展和增长，并将在多个方向上拓宽研究的可行性，例如，显著物体的感数 [46]，实例级显著性物体检测 [22]，基于弱监督的显着对象检测 [37]等等。然后，为了更深入地了解SOD问题，研究SOD算法的优缺点，并在不同的观点和要求下客观地评估模型性能，我们提出了一组属性（例如，外观变化）。最后，我们在SOC数据集上对现有SOD模型进行了基于属性的性能评估，评估的结果为未来的模型开发和模型比较开辟了充满希望的新方向。

致谢

本研究得到了国家自然科学基金（NO.61620106008,61572264），国家青年人才支持计划，天津市杰出青年学者自然科学基金（NO.17JCJQJC43700），华为创新研究计划的支持。

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: CVPR. pp. 1597–1604. IEEE (2009)
2. Alpert, S., Galun, M., Basri, R., Brandt, A.: Image Segmentation by Probabilistic Bottom-Up Aggregation and Cue Integration. In: CVPR. pp. 1–8. IEEE (2007)
3. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE TIP **24**(12), 5706–5722 (2015)
4. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE TPAMI **35**(1), 185–207 (2013)
5. Borji, A., Sihite, D.N., Itti, L.: Salient object detection: a benchmark. In: ECCV. pp. 414–429. Springer (2012)
6. Caesar, H., Uijlings, J., Ferrari, V.: COCO-Stuff: Thing and Stuff Classes in Context. In: CVPR. IEEE (2018)
7. Chen, T., Cheng, M.M., Tan, P., Shamir, A., Hu, S.M.: Sketch2photo: Internet image montage. ACM Transactions on Graphics (TOG) **28**(5), 124 (2009)
8. Chen, T., Lin, L., Liu, L., Luo, X., Li, X.: DISC: Deep image saliency computing via progressive representation learning. IEEE transactions on neural networks and learning systems **27**(6), 1135–1149 (2016)
9. Cheng, M.M., Hou, Q.B., Zhang, S.H., Rosin, P.L.: Intelligent Visual Media Processing: When Graphics Meets Vision. Journal of Computer Science and Technology **32**(1), 110–121 (2017)
10. Cheng, M.M., Mitra, N.J., Huang, X., Hu, S.M.: Salientshape: group saliency in image collections. The Visual Computer **30**(4), 443–453 (2014)
11. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.M.: Global contrast based salient region detection. IEEE TPAMI **37**(3), 569–582 (2015)
12. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results
13. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A New Way to Evaluate Foreground Maps. In: ICCV. pp. 4548–4557. IEEE (2017)
14. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment Measure for Binary Foreground Map Evaluation. In: International Joint Conference on Artificial Intelligence (IJCAI). pp. 698–704 (2018)
15. Gayoung, L., Yu-Wing, T., Junmo, K.: Deep Saliency with Encoded Low level Distance Map and High Level Features. In: CVPR. IEEE (2016)
16. He, J., Feng, J., Liu, X., Cheng, T., Lin, T.H., Chung, H., Chang, S.F.: Mobile product search with bag of hash bits and boundary reranking. In: CVPR. pp. 3005–3012. IEEE (2012)

17. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. IEEE TPAMI (2018)
18. Hou, Q., Dokania, P.K., Massiceti, D., Wei, Y., Cheng, M.M., Torr, P.H.S.: Bottom-Up Top-Down Cues for Weakly Supervised Semantic Segmentation. In: EMMCVPR. IEEE (2017)
19. Jiang, H., Cheng, M.M., Li, S.J., Borji, A., Wang, J.: Joint Salient Object Detection and Existence Prediction. Front. Comput. Sci (2018)
20. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations. In: MIT Technical Report (2012)
21. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using local affine regions. IEEE TPAMI **27**(8), 1265–1278 (2005)
22. Li, G., Xie, Y., Lin, L., Yu, Y.: Instance-Level Salient Object Segmentation. In: CVPR. pp. 247–256. IEEE (2017)
23. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: CVPR. pp. 5455–5463. IEEE (2015)
24. Li, G., Yu, Y.: Deep Contrast Learning for Salient Object Detection. In: CVPR. pp. 478–487. IEEE (2016)
25. Li, X., Zhao, L., Wei, L., Yang, M.H., Wu, F., Zhuang, Y., Ling, H., Wang, J.: DeepSaliency: Multi-task deep neural network model for salient object detection. IEEE TIP **25**(8), 3919–3930 (2016)
26. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: CVPR. pp. 280–287. IEEE (2014)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
28. Liu, N., Han, J.: DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection. In: CVPR. pp. 678–686. IEEE (2016)
29. Liu, T., Sun, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to Detect A Salient Object. In: CVPR. pp. 1–8. IEEE (2007)
30. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440. IEEE (2015)
31. Luo, Z., Mishra, A.K., Achkar, A., Eichel, J.A., Li, S., Jodoin, P.M.: Non-local deep features for salient object detection. In: CVPR. vol. 2, p. 7 (2017)
32. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV. vol. 2, pp. 416–423. IEEE (2001)
33. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR. pp. 724–732. IEEE (2016)

34. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
35. Wang, J., Jiang, H., Yuan, Z., Cheng, M.M., Hu, X., Zheng, N.: Salient Object Detection: A Discriminative Regional Feature Integration Approach. IJCV **123**(2), 251–268 (2017)
36. Wang, L., Lu, H., Ruan, X., Yang, M.H.: Deep networks for saliency detection via local estimation and global search. In: CVPR. pp. 3183–3192. IEEE (2015)
37. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR. pp. 136–145. IEEE (2017)
38. Wang, L., Wang, L., Lu, H., Zhang, P., Ruan, X.: Saliency detection with recurrent fully convolutional networks. In: ECCV. pp. 825–841. Springer (2016)
39. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. In: CVPR. IEEE (2017)
40. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Feng, J., Zhao, Y., Yan, S.: STC: A simple to complex framework for weakly-supervised semantic segmentation. IEEE TPAMI (2017)
41. Xia, C., Li, J., Chen, X., Zheng, A., Zhang, Y.: What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors. In: CVPR. IEEE (2017)
42. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV. pp. 1395–1403. IEEE (2015)
43. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: CVPR. pp. 1155–1162. IEEE (2013)
44. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR. pp. 3166–3173. IEEE (2013)
45. Zhang, G.X., Cheng, M.M., Hu, S.M., Martin, R.R.: A Shape-Preserving Approach to Image Resizing. Computer Graphics Forum **28**(7), 1897–1906 (2009)
46. Zhang, J., Ma, S., Sameki, M., Sclaroff, S., Betke, M., Lin, Z., Shen, X., Price, B., Mech, R.: Salient object subitizing. In: CVPR. pp. 4045–4054. IEEE (2015)
47. Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B., Mech, R.: Unconstrained salient object detection via proposal subset optimization. In: CVPR. pp. 5733–5742. IEEE (2016)
48. Zhang, J., Dai, Y., Porikli, F.: Deep Salient Object Detection by Integrating Multi-level Cues. In: Winter Conference on Applications of Computer Vision (WACV). pp. 1–10. IEEE (2017)

49. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: ICCV. pp. 202–211 (2017)
50. Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B.: Learning Uncertain Convolutional Features for Accurate Saliency Detection. In: ICCV. pp. 212–221 (2017)
51. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: CVPR. pp. 1265–1274. IEEE (2015)